DOCUMENT RESUME

ED 275 171 FL 016 075

AUTHOR Raatz, Ulrich
TITLE The Factorial Validity of C-Tests.
PUB DATE 84
NOTE 18p.; In: Practice and Problems in Language Testing. Papers from the International Symposium on Language Testing (7th, Colchester, England, 1984); see FL 016 066.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Attention Control; Cloze Procedure; Elementary Education; Factor Analysis; *Factor Structure; Foreign Countries; German; Grade 5; Grades (Scholastic); Higher Education; *Intelligence Tests; *Language Tests; Second Languages; *Test Validity
IDENTIFIERS *C Tests

ABSTRACT

This study describes the convergent, divergent, and factorial validity of a German C-Test in two parallel versions for German pupils in the fifth grade. The study investigated (1) the correlation between the C-Tests and the school grades in German, grammar, and English and with other language tests (convergent validity); (2) the correlation between C-Tests and intelligence and concentration tests (divergent validity); and (3) the factor structure produced by the tests used and the implications for the factorial validity of C-Tests. It is concluded from the results that despite the limitations of this pilot study, this examination of the C-Tests' factorial validity has produced results similar to those obtained on all previous validity studies of the C-Test, and that a close second-order relationship found between language proficiency and general intelligence is interesting, though not surprising. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED275171

THE FACTORIAL VALIDITY OF C-TESTS

ULRICH RAATZ
(COLCHESTER, OCTOBER 1983)

FL 016 075

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Cuthane

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE FACTORIAL VALIDITY OF C-TESTS

Ulrich Raatz (Colchester, October 1983)

1 Introduction

It was in Colchester at the Fourth International Language Testing Symposium of the IUS, almost exactly two years ago, that Kris Klein-Braley and I first presented the C-Principle and the earliest investigations with C-tests in English and German to an expert audience. Since then the interest in this type of test among language testers, psychologists and educationists has been steadily increasing. C-Tests have been used, or are being used, in Germany, England, Belgium and Finland. Rebecca Valette has spread the news in the United States of America. Kris was invited to Israel to talk to the Israeli Committee on Language Testing about C-Tests. Interested colleagues have taken up our challenge and developed C-Tests for their own target groups, for instance Edgar Süßmilch for the children of immigrant workers in Germany and Rüdiger Grotjahn for university students of French. First test forms have been developed for Spanish, Greek and Turkish. In Cologne at the Eurozentrum, where students from all over the world come to learn German, a C-Test is part of the regular test battery used for placement of incoming students, together with traditional discrete-point item tests.

In the last two years, Kris Klein-Braley and I have presented papers on the C-Test at around 12 national and international congresses. At this point of time we have conducted a total of 21 different empirical investigations into C-Tests and published the results. Other investigations are currently in progress. Yet others have been completed but have not yet been published, and it is the result of one of these investigations that I want to present in this paper.

In our research into C-Tests over the past two years we have been concerned with three main questions:

1. the examination of the reliability and validity of C-Tests for a variety of different groups;
2. investigations into the suitability of C-Tests for classification decisions;
3. examination of the linguistic and psychometric theory underlying C-Tests.

We have administered German and English tests to native speakers of these languages, to learners of English and German as a second language, to foreign language learners. Our subjects range from pupils in the third to tenth grades, and from the sixth form level to adults. In all, around 3 000 subjects have taken tests. All this research activity has not been able to discredit the C-Principle and the C-Tests.

2 Previous validation of C-Tests

In all our investigations into C-Tests up to now one of the most important questions has been that of validity. What do C-Tests measure?

Since C-Tests use several unsystematically selected texts, whose subject matter is as varied as possible, since in each case the mutilated words are a representative sample of all the words in the texts, and since, in addition, they are based on a linguistic theory, namely Oller's construct of pragmatic expectancy grammar, we claim that C-Tests are content valid tests.

We have used two different criteria to determine empirical validity. In 12 investigations into L1 groups we used teacher ratings as our criterion. This was either the last school grade given on a regular report or an estimation of current performance. Sometimes the global performance was evaluated, sometimes individual components such as orthography or grammar. In such cases the individual scores were averaged to obtain the criterion. Table 1 shows the distribution of the 12 validity coefficients.

Table 1

Distribution of validity coefficients for previously published investigations (Raatz and Klein-Braley 1983)

r_{tc}	f
0 - .10	1
.11 - .20	
.21 - .30	
.31 - .40	
.41 - .50	2
.51 - .60	2
.61 - .70	1
.71 - .80	4
.81 - .90	2
.91 -1.00	

In 11 investigations with LF groups we were able to make use of scores obtained from other test batteries as criteria. These tests were being used in placement procedures. Table 2 shows the results for these investigations.

Table 2

Results of validity studies for LF groups
(Raatz and Klein-Braley 1983; Grotjahn 1983)

	Duisburg	Köln	Bochum
Language	English	German	French
Population	German students of English	Foreign students	German students of French
Criterion	DELTA-Test 6 Subtests	Euro-Test 4 Subtests	Placement Test F 7 subtests
N of investigations	8	2	1
Validity	.77 - .91	.82/.85	.87

The results for the French C-Test are new. They were reported at the Congress of the German Association of Applied Linguistics by Rüdiger Grotjahn.

Factorial validity has been determined so far in 5 investigations. In the two studies performed at the Eurozentrum in Cologne a factor analysis was calculated over the C-Test and the four subtests of the Eurotest. In both cases a general factor emerged which explained around 70 % of the variance, and the C-Tests had their highest loadings on this factor. Similar results emerged in three investigations in Duisburg for students of English where factor analyses were performed over the C-Tests and the six subtests of the DELTA Test. However, since the number of variables involved in each case was relatively small, these results cannot be regarded as definitive.

We can sum up the results of these investigations by drawing the following conclusions:

1. In the classroom setting the C-Tests produce the same judgments as the experienced teacher in L1 teaching.

2. The C-Tests measure the same thing that complex language test batteries measure in LF-teaching.

However, the question of what the C-Test actually measures has still not been answered. The pilot study to be reported here is a first attempt to provide an answer to this question.

3 The purposes of this investigation

In this study we investigated the convergent, divergent and factorial validity of a German C-Test in two parallel versions for German pupils in the 5th grade. The individual questions we hoped to answer were as follows:

1. What is the correlation between the C-Tests and the school grades in German, grammar and English and with other language tests? (convergent validity).

2. What is the correlation between the C-Tests and intelligence and concentration tests? (divergent validity).

3. What factor structure is produced by the tests used, and what is the implication for the factorial validity of C-Tests?

In view of comments I have previously made (cf. Raatz 1981) about the use of factor analysis in attempting to prove theories of language I should like to point out explicitly that in this study factor analysis is used as a descriptive technique. No hypotheses about an expected factor structure had been set up in advance.

4 The sample

In Mülheim, a smaller town near Duisburg, a school centre has been established where the three traditional types of secondary schools in Germany share a communal campus. These school types are Hauptschulen, Realschulen and Gymnasien. Roughly speaking the Hauptschule is equivalent to the secondary modern school, the Gymnasium to the grammar school, and the Realschule lies somewhere in the middle. It is, however, much more common than the secondary technical school in England.

Our investigation was performed in one 5th grade form from each type of school. The results were then pooled for the analysis. We hoped (by pooling) to avoid the restriction of variance which would necessarily occur if we studied only pupils from one of the three types of secondary school, as compared to the variance in the 4th grade form, the last form of the primary school, where the pupils are still together.

The structure of the sample is shown in Table 3.

Table 3

The structure of the sample

	N	N boys	N girls	mean age	Number of non-German children
Hauptschule	18	7	11	11	4
Realschule	24	13	11	10,9	1
Gymnasium	33	20	13	10,8	1
all	75	40	35	10,9	6

A total sample size of N = 75 is rather a small group for a factor analysis. For this reason we also performed a cross-validation study, by examining the results obtained for the two halves of the sample separately. Even so the results of the study are more indicative than conclusive and should be interpreted with the necessary caution.

5 Variables

The variables listed in Table 4 were used in the study.

Table 4

List of the variables included in the study

1. German C-Test in 2 parallel versions A and B
2. School grade in German
3. Teacher rating in grammar
4. School grade in English
5. School grade in mathematics
6. Diagnostic German Test (DTD 4-6) with the subtests
   (PW) PV Passive Vocabulary
   (AF) FA Finding Analogies
   (TS) TS Text Structuring
   (IV) IC Instruction Comprehension
   (LV) RC Reading Comprehension
   (AW) AV Active Vocabulary
   (in brackets the German abbreviations)
7. Orthography test DRT 4-5
8. Non-verbal intelligence test FDA 3-6
9. Concentration test d2

The two parallel forms of the C-Test (A and B) contain 6 texts each, and each text has 20 mutilated words. Eight texts had already been tried out in test forms, 4 texts were new. Administration of the C-Test took 40 minutes.

The school grades for German, English and mathematics were those given at the end of the last half-year. In Germany school grades can, in principle, vary from 1 (very good) to 6 (beyond the pale), and they are normed on the performance of the individual class. In the calculations the poles were reversed in order to avoid negative correlations or factor loadings.

Grades for grammar are not envisaged in the normal school grading scheme. However, here too the teachers used the familiar school grades as rating categories.

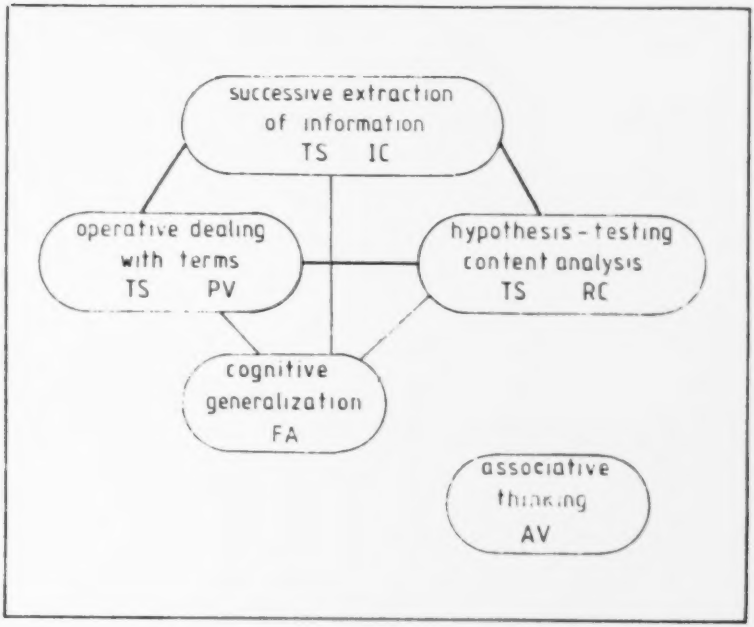
The Diagnostic German Test (DTD 4-6) was developed in 1980 by NAUCK and OTTE. This test takes as its model the assumption that all achievements in German mother tongue teaching can be attributed to 6 language-bound fundamentals, which are operationalised in the 6 subtests of the DTD listed and described below.

1. PV Passive vocabulary. Here the pupil is expected to find familiar words which are distorted in some way. To do this he must use his passive vocabulary. (22 items, 8 minutes.)
2. FA Finding analogies. Here comparisons, classifications and relations are given and analogies must be discovered. (22 items, 8 minutes.)
3. TS Text structuring. Simple sentences whose words are printed without spacing between them must be structured by inserting vertical lines. (22 items, 10 minutes)
4. IC Instruction comprehension. Here instructions of the type often found in instruction sheets or school workbooks have to be followed. The difficulty in processing lies in the connection of simple actions by complex operators. (17 items, 12 minutes.)
5. RC Reading comprehension. Two texts are given (mini-stories) with a number of 4-choice multiple-choice questions. Choice of the correct, or the best response is determined by the ability to understand the text without help. (18 items, 16 minutes.)
6. AV Active vocabulary. The pupil has to find as many words as he can which begin with the letter given. This indicator for word fluency is primarily dependent on the range of vocabulary available to the individual since the time limits are generous. (2 attempts, 2 minutes each.)

The authors combine one or two of these fundamentals to 5 part constructs, which form a model structure of language learning, as shown in Fig. 1.

Figure 1

The structure of part constructs in language learning
(Nauck and OTTE, 1980)



The three constructs at the top, which all contain the fundamental TS (text structuring), have the greatest importance in this structure. The DTD 4-6 test demands the ability to perform pragmatic operations and analyses on language; it does not demand the ability to describe the functions and working of language in a cognitive fashion. For this test two parallel forms are available. The total score for the test and the result for each individual subtest were entered into the analysis as a raw score.

The orthography test DRT 4-5 was developed by Meis (1969). The test uses a format which we in Germany call the Lückendiktat. The subject receives the text in a form similar to the rational deletion cloze test. However in this type of test the text is read to him and he inserts the critical words into the blanks. The score is the sum of the correctly written words, and was entered into the analysis as a raw score.

BEST COPY AVAILABLE

In order to measure intelligence I selected the non-verbal intelligence test FDA 3-6, which is also available in two parallel forms (Wendeler 1973). This test measures chiefly the general intelligence factor, together with the space factor. It has 3 subtests with 21,21 and 20 items respectively and each subtest has an time-allowance of 15 minutes.

In the first subtest a sequence must be continued using one of the 4 alternatives given. In the second subtest the figure which does not fit must be discovered. In the third subtest the missing figure in a matrix must be restored. Here again the raw score was used for analysis.

The concentration test d2 (also called the attention-stress test) was developed by Brickenkamp (1972). The subjects receive a test sheet with 14 lines of text; each line has 47 characters (d, p). The subject must cross out every d which has two dashes with it. He has 20 seconds for each line. The final score is the number of characters attempted minus the number of errors.

6 Test administration

The test sessions took place in a three week period during April 1983. For each group several sessions were necessary; in each case the first lesson periods were used.

7 Results

7.1 C-Test

Table 5 shows the means, difficulty values (P), standard deviations and reliabilities for the parallel forms of the C-Test.

Table 5

Means, standard deviations, difficulty levels, and reliability coefficients for the parallel C-Test forms

	test form	
	A	B
N	37	38
$\bar{x}$	60.2	67.4
S	20.3	19.7
P	50.2	56.2
r_{tt}	.78	.81

In view of the fact that both test versions were put together arbitrarily, that 4 new texts were used, and that no item analysis had been performed beforehand, the reliability coefficients are very satisfactory.

Both test forms have a medium level of difficulty for the group overall. The results are, however, not directly comparable because of the difference of 7 points between the means. For this reason a T-transformation was performed on both sets of scores, and these T-values were used for all analyses.

7.2 Equating of the school grades

It is customary in German to use an internal class-scale when determining school grades. Pupil performance is evaluated against the mean performance of the entire class, not against a criterion or some other absolute value. For this reason grades given in different classes are not comparable if the classes differ in their average performance level. Whether this is the case can be examined for instance by administering a test.

In this study precisely these conditions obtained: There were differences in performance level between the three classes, since the conception behind the German tripartite school system ist exactly that children with different types of ability (and thus, by implication, ability levels) should go to the different types of school.

For this reason I decided to choose the variable which correlated most highly with the school grades as an anchor, and to determine correction factors for the grades by regression procedures. The variable chosen was the total score on the DTD test. Table 6 shows the correction terms for the individual variables and school types.

Table 6

Correction terms for the grades

	German grammar	English	Maths
Gymnasium	0	0	0
Realschule	0,4	0,5	0,6	0,6
Hauptschule	0,7	0,9	1,0	1,1

The corrected school grades were used for further analysis. The total score for DTD was ignored in subsequent calculations in order to avoid artefacts.

7.3 The correlation coefficients

In order to examine convergent and divergent validity for the C-Tests the correlations were calculated between the C-Test scores and all the other variables, both for the entire group and for the subgroups which took the different C-Test forms A and B. Table 7 shows the correlations with the individual variables ranked in order of magnitude.

Table 7

Correlations between C-Tests and the other variables

all		A		B	
DRT	.68	DRT	.63	TS	.76
TS	.66	PV	.62	DRT	.75
G German	.61	TS	.57	G German	.65
G gramm	.60	G German	.57	G gramm	.64
G Maths	.57	G gramm	.57	G Maths	.64
G Engl	.54	G Engl	.57	FDA	.57
IC	.54	IC	.54	G Engl	.54
FDA	.51	G Maths	.49	IC	.53
PV	.49	FDA	.46	RC	.50
FA	.44	FA	.42	FA	.46
RC	.43	AV	.39	AV	.40
AV	.40	RC	.36	PV	.36
D2	.20	D2	.25	D2	.14

r_{SAB} = 0,63 without PV r'_{SAB} = 0,87

We can draw the following conclusions from Table 7:

1. For the entire group the C-Test correlates most highly with the orthography test, with the grades for German and grammar, and with the subtest TS, which measures the ability to structure written linguistic material (= texts!). According to the authors of the DTD, TS reflects "a global performance which is necessary in different areas of language acquisition and mother tongue language proficiency". Here the correlations all reach at least the .6 level.
2. We find medium correlations (between .4 and .6) between C-Test scores and the school grade in English, the vocabulary and reading comprehension tests, and the tests which measure general intelligence and other intelligence factors (space, analogies, instruction comprehension).
3. There is no relationship to the concentration test.
4. The two subtests TS and IC in the upper part of the table constitute the part construct "successive extraction of information" according to the authors of the DTD test (see figure 1).

A comparison of the columns A and B in Table 7 shows very similar structures for each of them. An exception is the subtest Passive Vocabulary which correlates almost twice as high with the A form of the C-Test as with the B form.

If we calculate a rank correlation over the criterion variables between the two colums a value of .63 emerges if PV is included, and of .87 if this subtest is omitted.

A cautious interpretation of these results would be the following:

1. C-Tests measure chiefly general language proficiency, which is operationalised in the teacher ratings, the orthography test and the construct of successive extraction of information.
2. C-Tests also measure, though to a lesser degree, certain aspects of intelligence.
3. C-Test performance is independent of the ability to concentrate or of speed factors.
4. These results were confirmed in the cross-validation; The parallel test versions A and B are equivalent in their performance.

7.4 The factor analysis

In order to determine the factorial validity of the C-Tests intercorrelations between the variables both for the entire group and for each subgroup taking each of the two C-Test forms were calculated separately. The three factor analyses used the common factor method with iterative estimation of the communalities and a varimax rotation.

Factor extraction was broken off after the second factor, in spite of the fact that there were further factors with eigenvalues greater than 1 because:
- the proportion of variance explained diminished sharply
- the sample group was relatively small
- only two factors proved to be stable in the cross-validation study.

For the entire group the first two factors explained 61 % of the variance, for each of the subgroups 65 % and 61 % respectively.

Table 8 shows the loadings in factors 1 and 2 for the entire group after rotation. In order to bring out the factor structure more clearly those loadings have been marked which reach .6 or higher. The variables have been ordered according to the magnitudes of their loadings.

Table 8

Factor matrix for the entire group

variable	factor I	factor II
DRT	.87 o	-.04
TS	.80 o	.22
G gramm	.77 o	.32
G Engl	.74 o	.25
G German	.73 o	.27
C-Test	.71 o	.33
d2	-.11	.74 o
G Maths	.58	.65 o
FDA	.48	.63 o
IC	.43	.63 o
FA	.39	.61 o
RC	.49	.20
PV	.46	.26
AV	.23	.39

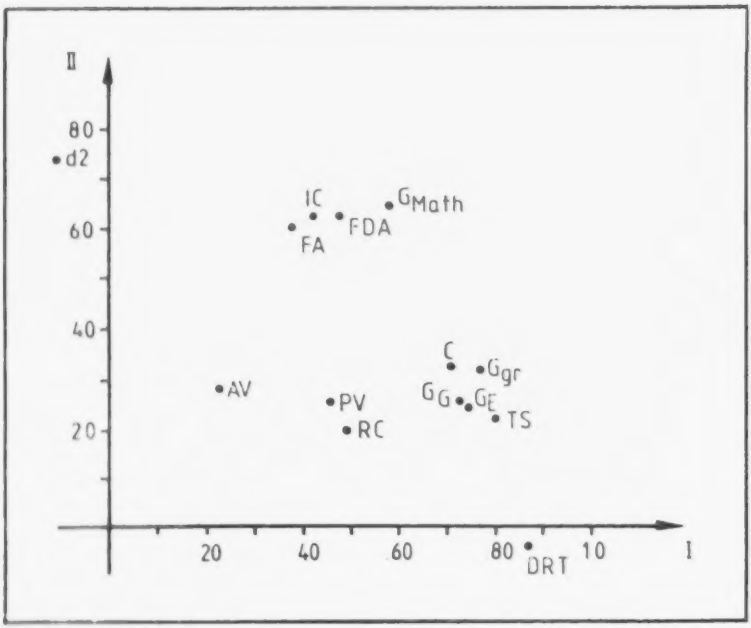
The first factor may be interpreted as a general language proficiency factor. This is supported by the fact that the dictation, which is a typical integrative language test from Spolsky's post-modern trend (Spolsky 1981), also loads high on this factor together with the school grades (also global measurements of language proficiency) and the subtest TS which was also described by the test authors as "global language performance". And the C-Test too has high loadings on the first factor.

The second factor has high loadings in the non-verbal intelligence test, in the maths grade, in the concentration test (which had a high speed component), and in the subtests instruction comprehension and analogies, which are, of course, subtests which are often found in other intelligence tests. This factor can be interpreted as a speeded factor of logical thinking. The C-Test has low loadings on this factor.

The factor matrix shows that it was not possible to achieve simple structure. This emerges very clearly in Figure 2, which is a graphic representation of the factor structure.

Figure 2

Factor structure for the entire group



Obviously the only way to achieve an ideal simple structure would be to select an oblique rotation. If the factor axes are permitted to be non-orthogonal, however, they have fairly high correlations with each other.

For the cross validation study a factor analysis of the same type was performed for each of the subgroups taking the parallel C-Tests A and B. In Table 9 the factor matrices for the two groups are compared with each other. Here, only loadings of .6 and higher have been included.

Table 9

Factor matrices in the two subgroups

	A		B	
Variable	I	II	I	II
DRT	.92		.75	
TS	.87		.70	
G gramm	.80		.75	
G Engl	.73		.81	
G German	.66		.84	
C-Test	.66		.72	
d2		.71		(.45)
G Maths	(.43)	.72	.69	.62
FDA		.74		.60
IC		(.51)		.86
FA		.78		.60
RC				
PV				
AV		.64		(-.02)

If we take the small number of subjects into consideration, the similarity is remarkable. The only marked differences are

- in group B the maths grade has high loadings on both factors
- the subtest active vocabulary has a high loading on the intelligence factor for group A, but for group B it has no loading on either factor.

In view of the sample sizes these differences should not be interpreted.

8 Interpretation

This investigation is a pilot study, and therefore intended to be exploratory rather than conclusive. There are three reasons why the results must be interpreted with caution.

1. The sample size was small, particularly in the two subgroups.
2. As in any factor analytic study there is always the danger (cf. Raatz 1981) that the results are predetermined by the variables selected for inclusion. For instance it is quite within the bounds of possibility that no separate intelligence factor would have emerged if I had chosen to use a test of verbal intelligence, or if the concentration test had been omitted.

3. As a result of the equating of grades the true variance was increased, but very probably the estimation of the correction terms also increased the error variance. It is not possible to evaluate the effects of this on the results obtained.

However, provided we bear all these critical points in mind, we can still conclude that the examination of the factorial validity of C-Tests has produced similar results to all the validity studies performed on the C-Tests so far:
German C-Tests for L1-learners - native speakers of the language - measure primarily general language competence.

The close second order relationship between language proficiency and general intelligence, as shown in Figure 2, is very interesting, but really not surprising. There are empirical studies in psychology and linguistics with similar results, and sophisticated theories to explain them.

At this point, I shall only mention Oller (1978), who states: "language proficiency ... may account for the lion's share of variance in so-called IQ tests".

But the examination of this relationship was not the purpose of our study. Our aim was to investigate the validity of the C-Test, and here we have been relatively successful.

We plan to repeat the study with a larger sample taken from two different school grades. If this replication confirms the results presented here, then we will have very strong evidence for the validity of the C-Test. A further question in this planned investigation will be the relationship between C-Tests and IQ tests.

Brickenkamp, R. 1972. Test d2. Aufmerksamkeits-Belastungs-Test. Göttingen: Hogrefe.

Grotjahn, R. 1983. Entwicklung und Einsatz eines C-Tests "Französisch". Paper presented at the 14th Conference of the GAL, Duisburg, Sept. 1983.

Meis, R. 1969. DRT 4-5 Diagnostischer Rechtschreibtest für 4. und 5. Klassen. Weinheim, Berlin, Basel: Beltz.

Nauk, J. und Otte, Rolf. 1980. DTD Diagnostischer Test Deutsch. Braunschweig: Westermann.

Oller, J.W., Jr., 1978. "How important is language proficiency to IQ and other educational tests?". In Oller, J.W., Jr. and Perkins, K. (eds). 1978. Language in Education. Rowley, Mass.: Newbury House. 1-15.

Raatz, U. 1981. Sprachtheorie und Faktorenanalyse. Duisburg und Essen: Interuniversitäre Sprachtestgruppe IUS Workpapers 9.

Raatz, U. und Klein-Braley, C. 1983. "Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test: Theorie und Praxis". In Horn, R., Ingenkamp, K., Jäger, R.S. (Hrsg). 1983. Tests und Trends 3. Weinheim und Basel: Beltz. 107-138.

Spolsky, B. 1981. "Some ethical questions about language testing". In Klein-Braley, C. & Stevenson, D. (Eds). 1981. Practice and problems in language testing I. Frankfurt/M. und Bern: Lang. 5-21.

Wendeler, J. 1973. FDA 3-6 Frankfurter Denkaufgaben. Weinheim: Beltz.

The tests were supervised by Peter Klatte who performed the investigation as part of his examination dissertation. I should like to take this opportunity of thanking him.